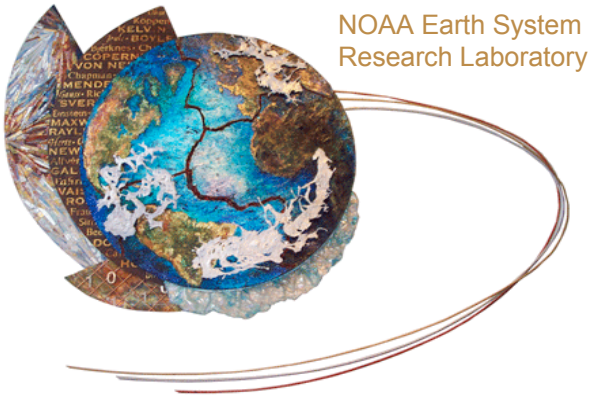


Calibrating and Combining Ensemble Predictions

Tom Hamill

NOAA Earth System Research Lab

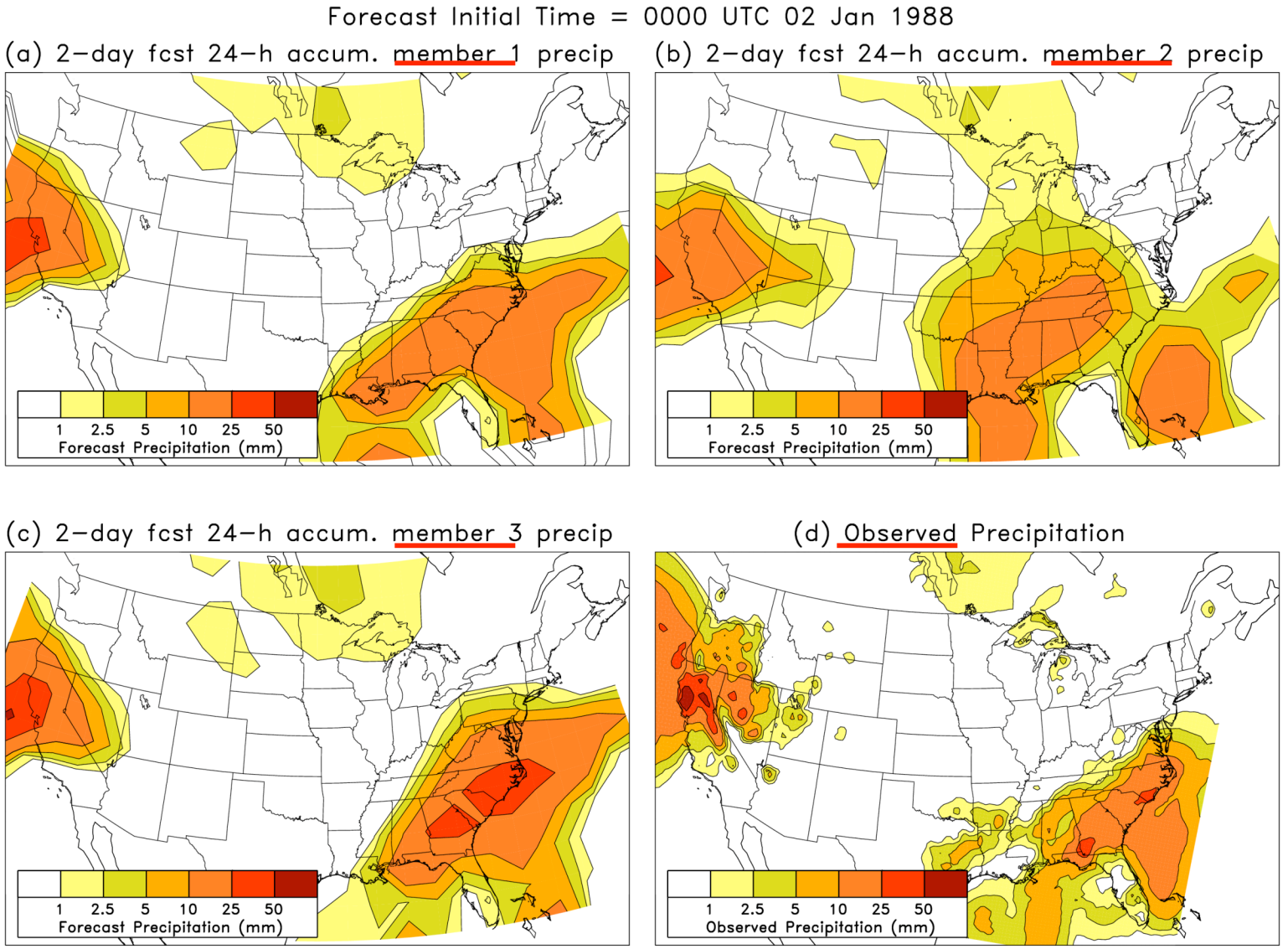
tom.hamill@noaa.gov

Definitions

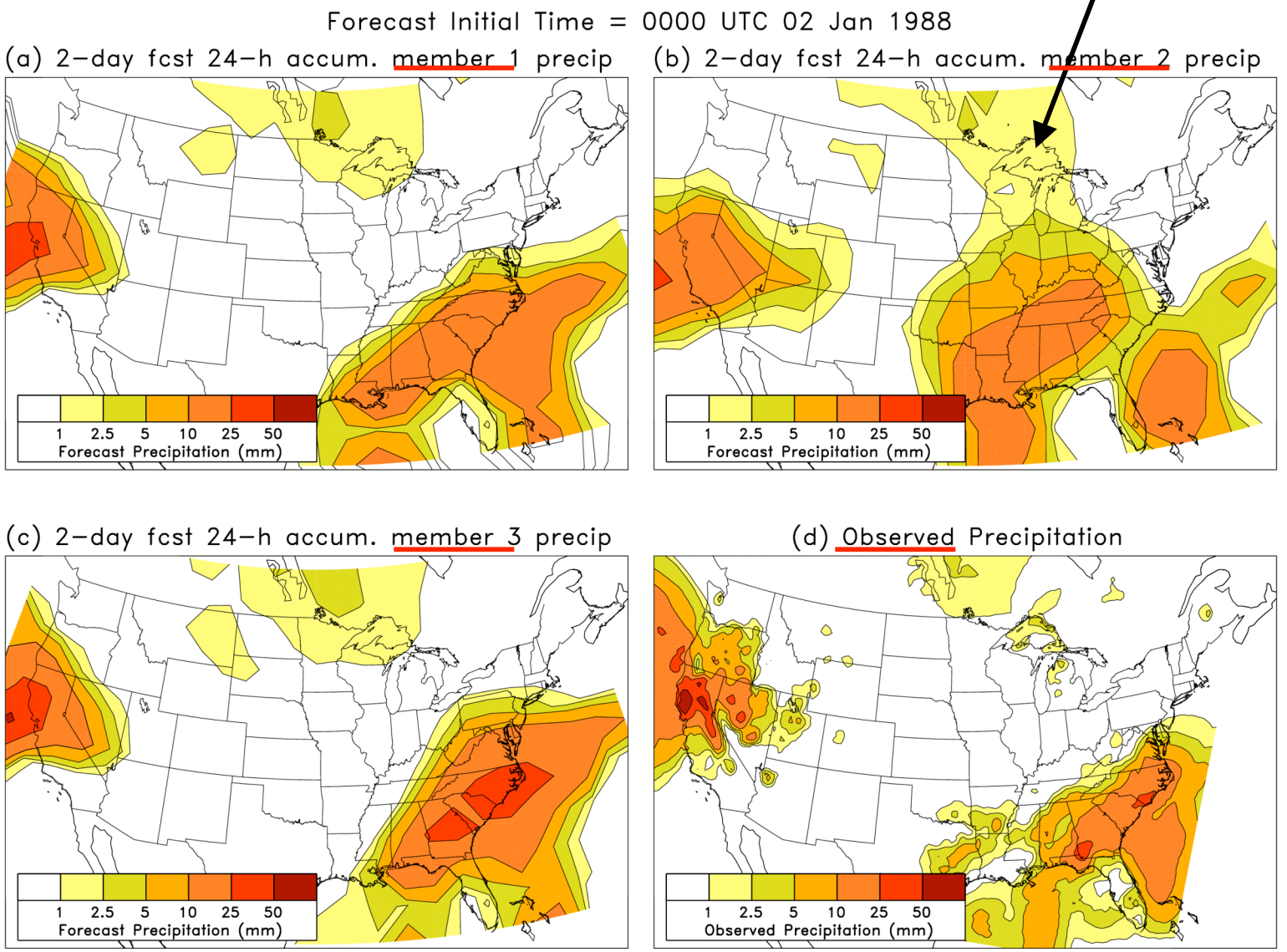
- **Calibration (part 1)**: $f\left(\mathbf{x}^t \middle| \mathbf{x}^f\right)$; the statistical adjustment of the (ensemble) forecast
  - Rationale 1: Infer large-sample probabilities from small ensemble.
  - Rationale 2: Remove bias, increase forecast reliability while preserving as much sharpness as possible. Guided by discrepancies between past observations and forecasts.

- **Combination (part 2)**: the formation of probability estimates using ensembles or control runs from multiple sources. May involve calibration.

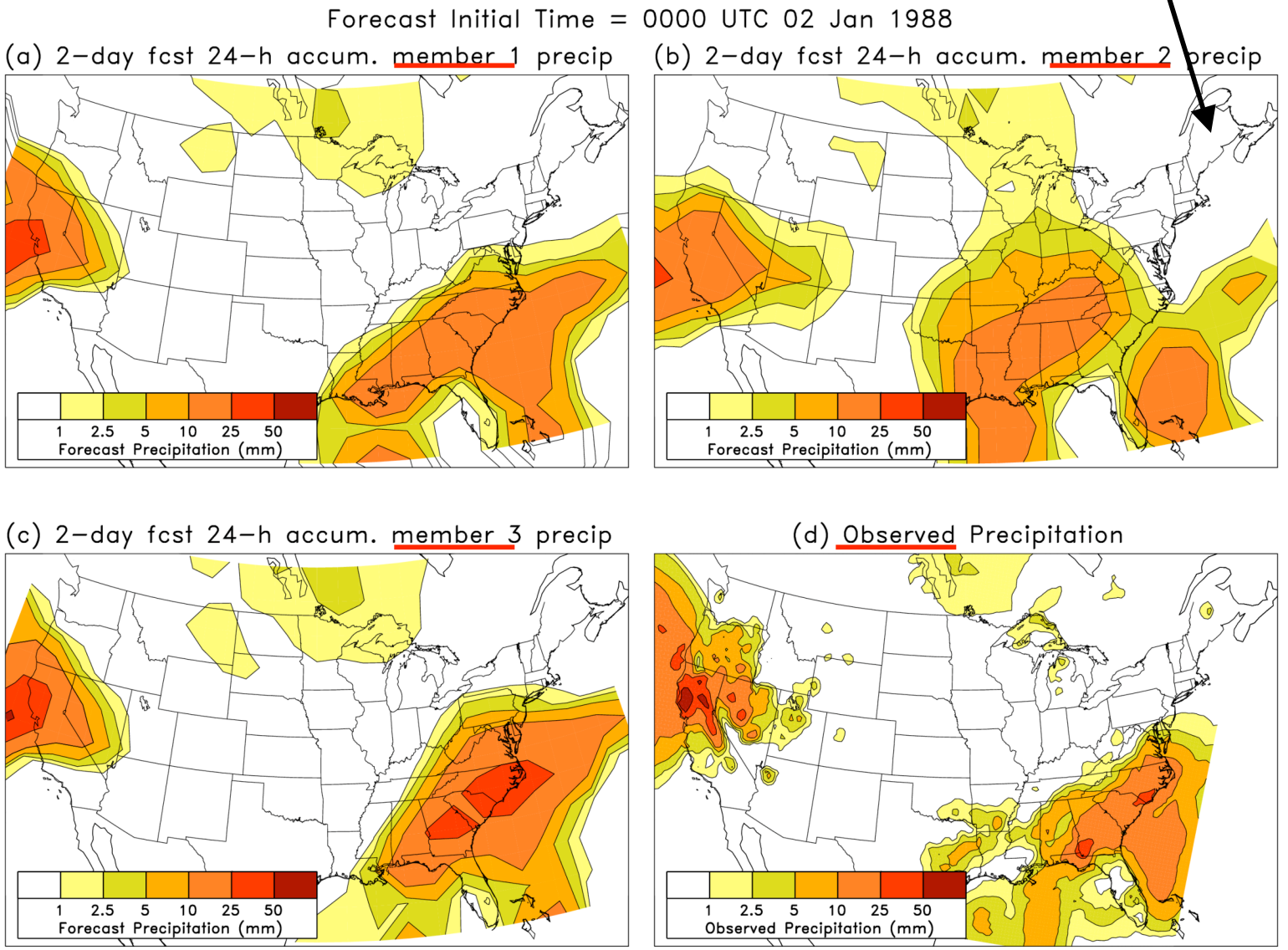
Ensemble-base probabilistic forecasts: problems we'd like to correct through calibration



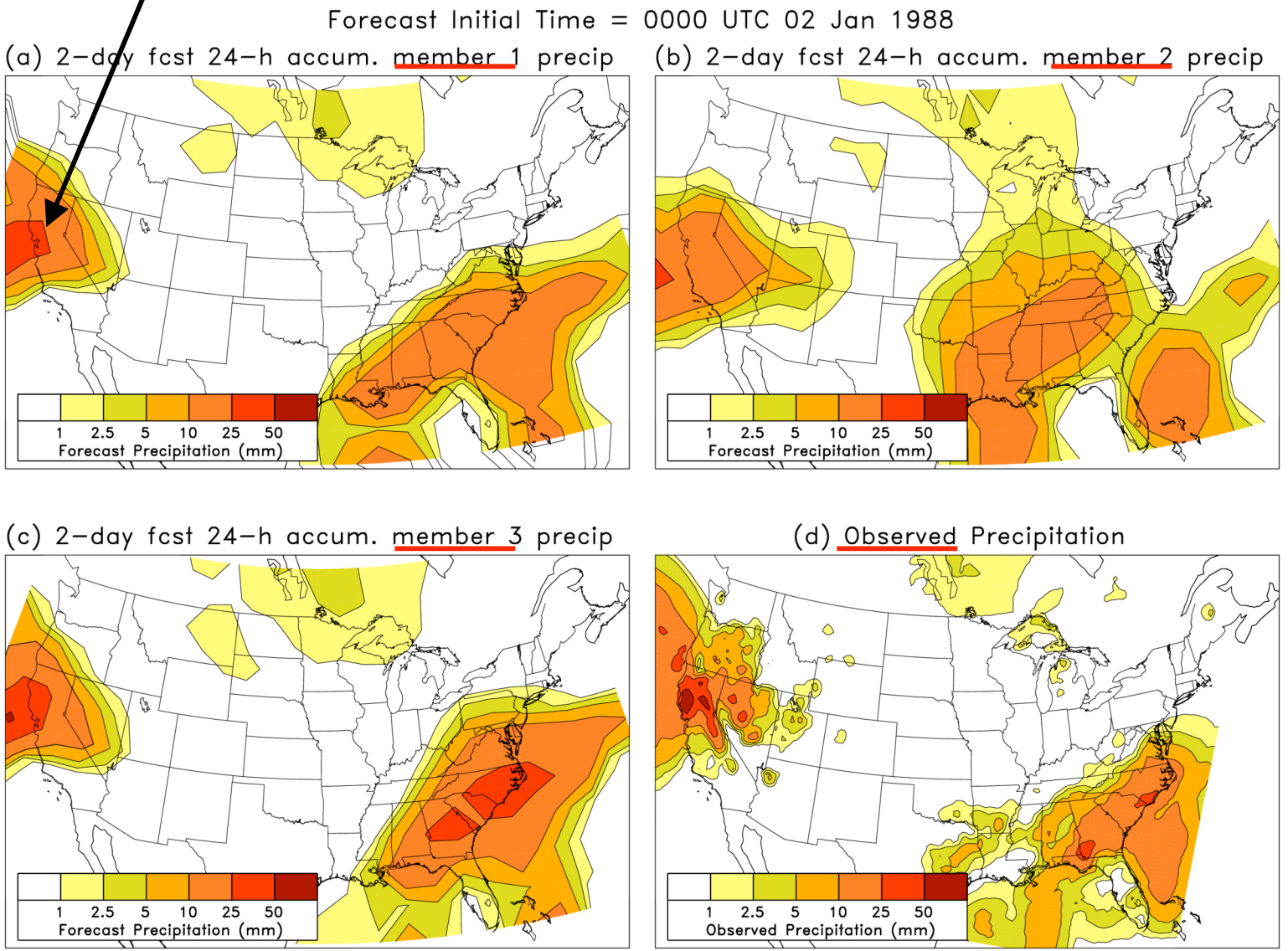
(1) bias (drizzle over-forecast)



(2) ensemble members too similar to each other.



(3) Ensembles are too smooth, not capturing intense local precipitation due to orographic forcing. *Downscaling* needed.



Calibration questions

- Is there a best technique, or best for this particular forecast problem? Different techniques may be needed for:
  - Errors are ~normally distributed, ~stationary, vs.
  - Distributions with long tails
- How much training data (past forecasts & observations) do you have / need?
  - More needed to do good job with rare events.
  - Lots more work involved in trying to get a good result with a short training data set.

Disadvantages to calibration

- Calibration won't correct the underlying problem. Prefer to achieve unbiased, reliable forecasts by doing numerical modeling correctly in the first place.
- No one general approach that works best for all applications.
- Corrections may be model-specific; the calibrations for NCEP v 2.0 may not be useful for ECMWF, or even NCEP v 3.0.
- Could constrain model development. Calibration ideally based on long database of prior forecasts (reforecasts, or hindcasts) from same model. Upgrading model good for improving raw forecasts, may be bad for skill of post-processed forecasts.
- Users beware: Several calibration techniques that have been recently proposed are conceptually flawed / only work properly in certain circumstances.

Calibration review

- Adjusting for sample size, no model-error correction
- Simple methods
  - Gross bias correction
  - Linear regression
  - Kalman filters
- More complex methods
  - Logistic regression
  - Rank histogram-based calibration
  - Dressing
  - Bayesian model averaging
  - CDF corrections
  - Non-homogeneous Gaussian regression

Inferring large-sample pdf from small ensemble: fitting parametric distributions

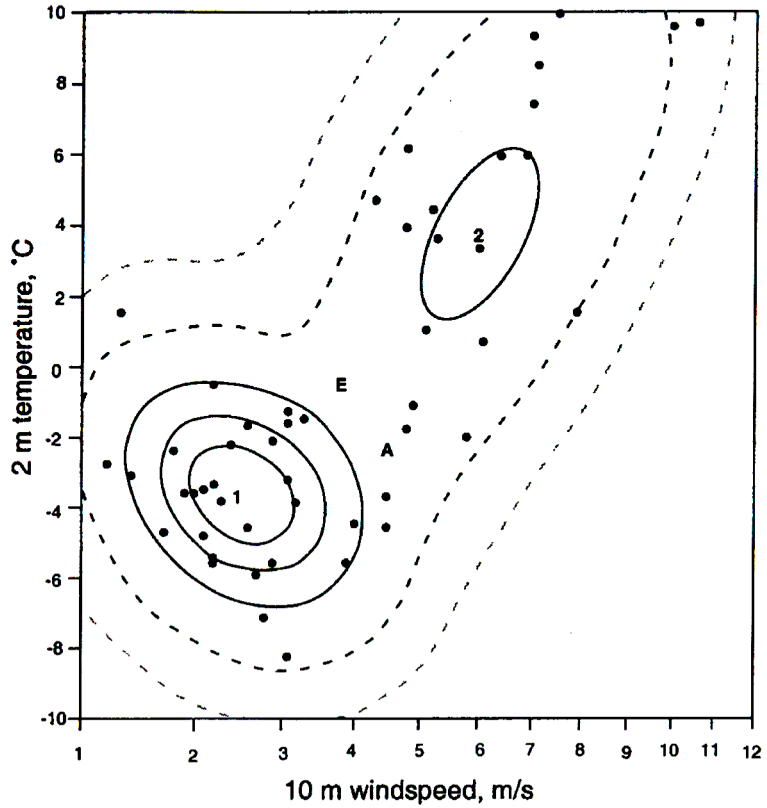
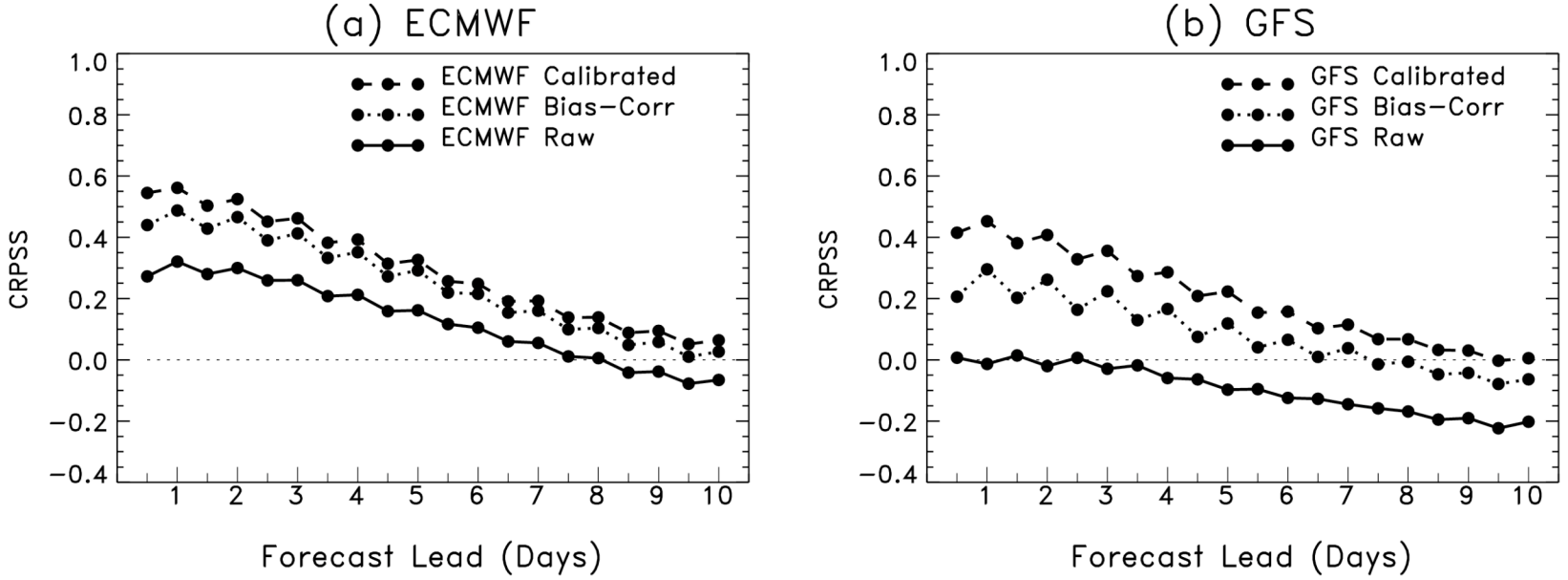


Figure 2. Example ensemble distribution with fitted Gaussian mixture, jointly for the temperature and wind-speed forecast at 12 UTC 10 January 1997 at Manchester, made at the 180 h lead time. Dots indicate individual forecasts made by the 51 ensemble members, with the ensemble mean located at 'E'. The two bivariate Gaussian densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are centred at '1' and '2', respectively, and the smooth lines indicate level curves of their mixture $f_{\text{mix}}(\mathbf{x})$, formed with $\alpha = 0.57$ (see text). Contour interval is 0.05, and the thick and thin dashed lines are for 0.01 and 0.001, respectively. Subsequent verifying analysis is 'A'.

Wilks (*QJRMS*, **128**, p 2821) explored fitting parametric distributions, or mixtures thereof, to ECMWF forecasts in perfect-model context. Power-transformed non - Gaussian variables prior to fitting. Goal was smooth pdfs, not bias/spread corrections.

Gross bias correction

- Given sample of past forecasts $x_1, \ldots, x_n$ and observations $y_1, \ldots, y_n$, gross bias correction is simply $\overline{y} - \overline{x}$

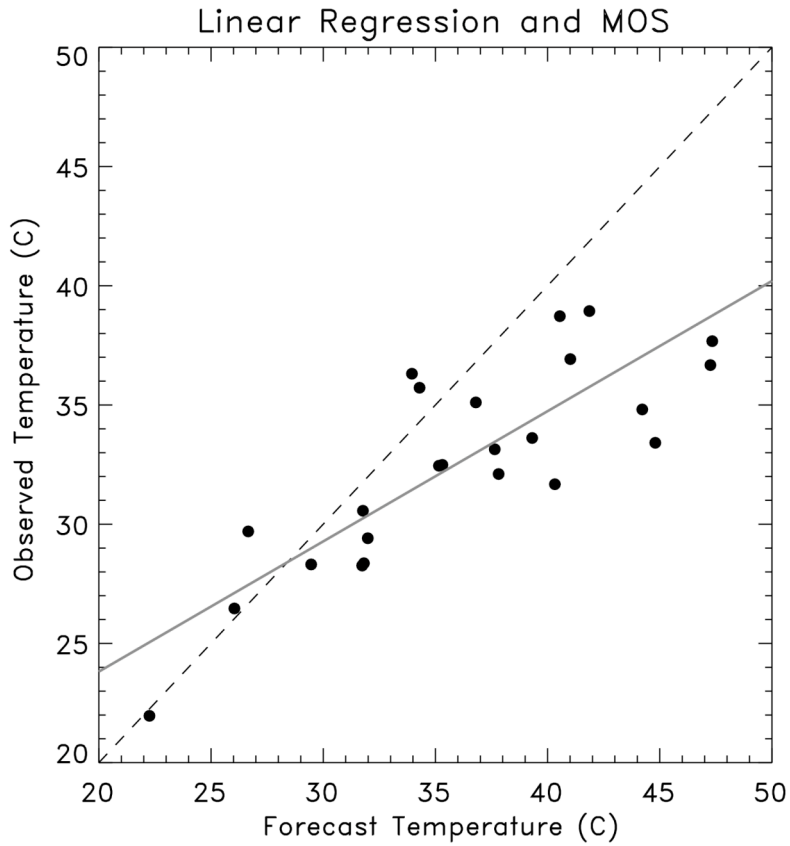


In surface-temperature calibration experiments with NCEP's GFS and ECMWF, simple gross bias correction achieved a large percentage of the improvement that was achieved through more sophisticated, bias+spread correction.

Ref: Hagedorn et al., MWR, 2008, in press.

Linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Corrects for bias; when no skill, regresses to sample climatology.

Diagnostics include statistics on error, so can infer pdf.

Multiple linear regression, with multiple predictors, often used.

Ref: any applied statistics textbook

Model Output Statistics (“MOS”)

many elements based on multiple linear regression

```
KBID   GFS MOS GUIDANCE    2/16/2005  1800 UTC
 DT /FEB  17                   /FEB  18                 /FEB  19
 HR   00 03 06 09 12 15 18 21 00 03 06 09 12 15 18 21 00 03 06 12 18
 N/X              32          40          25          35       19
 TMP  42 39 36 33 32 36 38 37 35 33 30 28 27 30 32 31 28 25 23 19 27
 DPT  34 29 26 22 19 18 17 17 17 17 17 15 14 13 11  8  7  6  5  2  4
 CLD  OV FW CL CL SC BK BK BK BK BK BK BK SC BK BK BK BK FW CL CL CL
 WDR  26 30 32 32 32 31 29 28 30 32 31 31 31 31 30 29 31 32 33 33 27
 WSP  12 12 12 11 08 08 09 08 09 09 10 10 10 12 13 13 15 16 15 09 08
 P06        17     0     0     0     4     0    10     6     8  0  0
 P12              17           0          10          17        8
 Q06         0     0     0     0     0     0     0     0     0  0  0
 Q12              0           0          0           0        0
 T06       0/ 2  0/ 0  1/ 0  1/ 2  0/ 1  0/ 1  1/ 0  0/ 1  0/ 0  0/ 0
 T12                 1/ 0        1/ 2        1/ 1        0/ 1  0/ 0
 POZ   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 POS  13 47 70 84 91100 96100100100100 92100 98100100100 94 92100100
 TYP   R  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S
 SNW                                     0                    0
 CIG   7  8  8  8  8  8  8  8  8  7  7  7  8  7  7  7  8  8  8  8  8
 VIS   7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7
 OBV   N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N
```

US: Statistical corrections to operational US NWS models, some fixed (NGM), some not (Eta, GFS). Refs: http://www.nws.noaa.gov/mdl/synop/index.htm, Carter et al., *WAF*, **4**, p 401, Glahn and Lowry, *JAM*, **11**, p 1580. **Canadian** models discussed in Wilson and Vallee, *WAF*, **17**, p. 206, and *WAF*, **18**, p 288. **Britain**: Met Office uses “updateable MOS” much like perfect prog.

Kalman filter

Today's forecast bias estimate

Yesterday's bias estimate

Yesterday's observed bias

$$\hat{b}_t^f = \hat{b}_{t-1}^f + K_t\left(\varepsilon_t - \hat{b}_{t-1}^f\right)$$

Kalman gain: weighting applied to residual

Pro:

- memory in system, amount tunable through K_t
- adaptive

Con:

- takes time to adapt after regime change

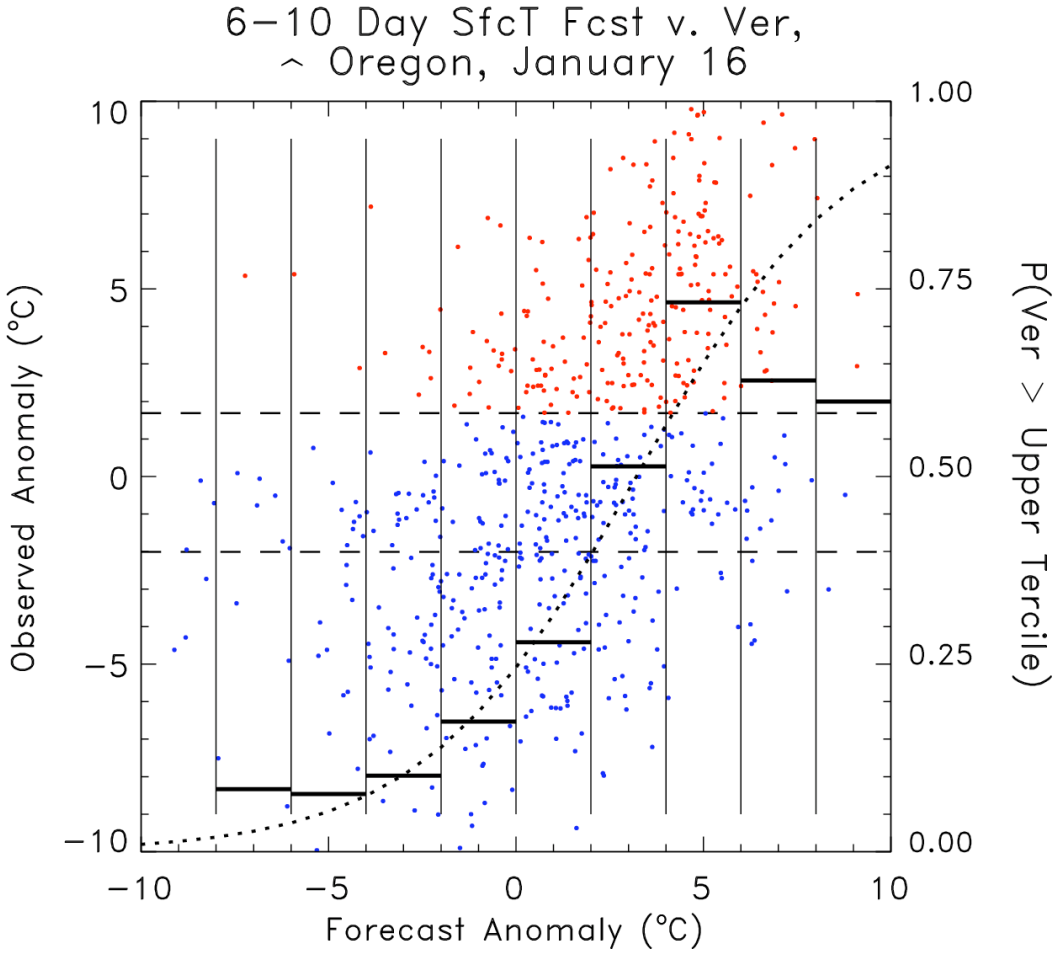
Ref: Cheng and Steenburg, conferences.dri.edu/WxPrediction/Weather12/Cheng_Steenburgh.ppt

Logistic regression

- For each grid point (or station) let *x* = continuous predictor data (ens. mean forecast value), *y* = binary predictand data (1.0 if predicted event happened, 0.0 if not).
- Problem: Compute P(*y* =1.0 | *x*) as a continuous function of *x*.
- Logistic Regression: $$P = 1 - \frac{1}{1 + exp\ (\beta_0 + \beta_1 x)}$$

Ref: any applied statistics text.

Logistic regression using a long data set of observed and forecast anomalies

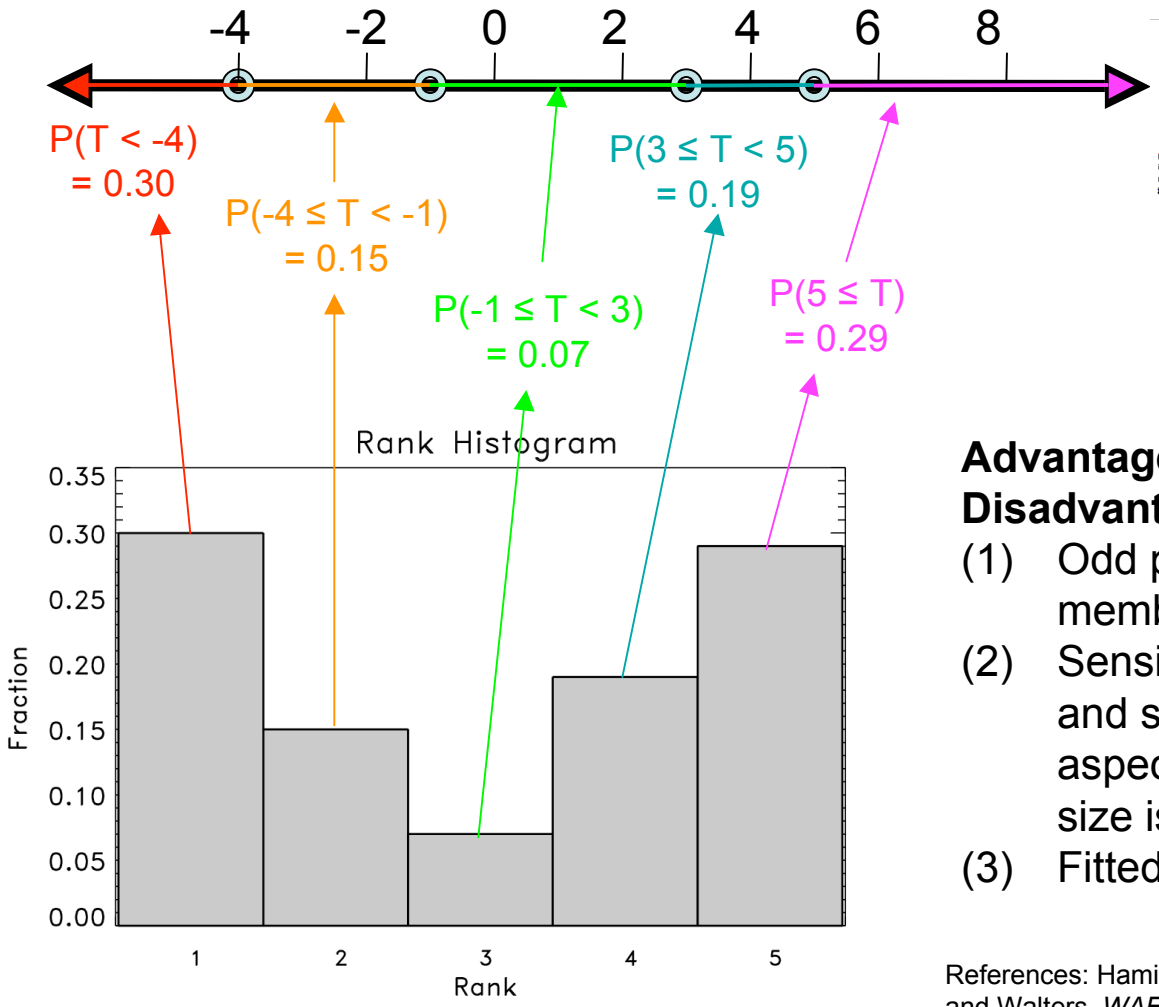


Seeking to predict probability of warmer than normal conditions (upper tercile of observed). Using reforecasts (a later talk), we have 23 years of data. Let's use old data in a 31-day window around the date of interest to make statistical corrections.

Dashed lines: tercile boundaries
Red points: samples above upper tercile
Blue points: samples below upper tercile
Solid bars: probabilities by bin count
Dotted line: logistic regression curve

Ref: Hamill et al. MWR, June 2004

Ensemble calibration: rank histogram techniques



NCEP MRF precipitation forecasts, from Eckel and Walters, 1998

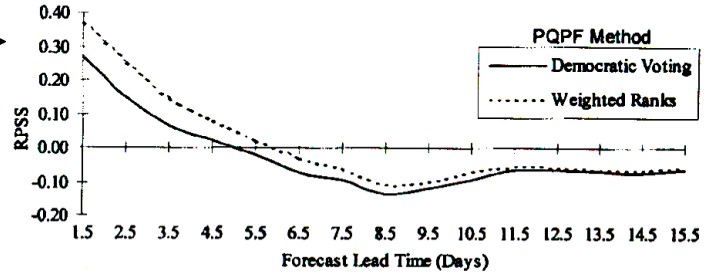


FIG. 10. Ranked probability skill score (RPSS) results for all forecast lead times.

Advantages: Demonstrated skill gain

Disadvantages:

(1) Odd pdfs, especially when two ensemble members close in value.

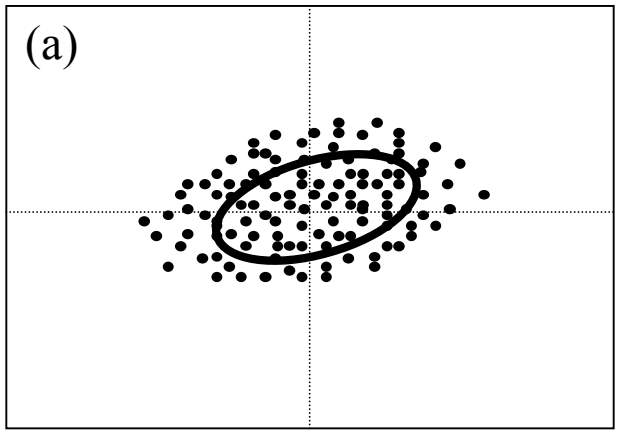
(2) Sensitive to shape of rank histogram, and shape of histogram may vary with aspects like precip amount --> sample size issues.

(3) Fitted parametric distributions as skillful

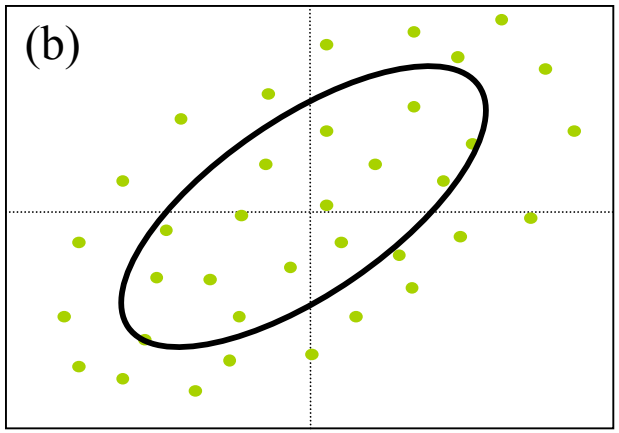
References: Hamill and Colucci (*MWR*, 1997, 1998; Eckel and Walters, *WAF*, 1998; used at Met Office)

Dressing methods

Original Ensemble



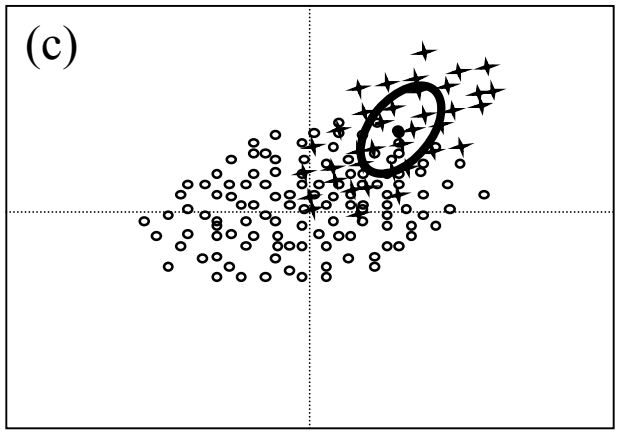
Cov(ens mean errors)



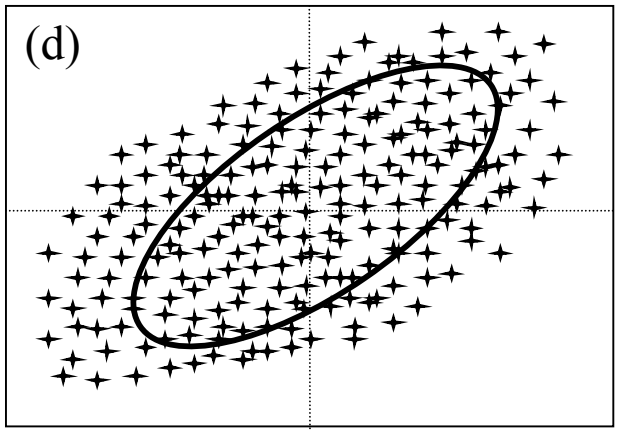
Method of correcting spread problems. Assume prior bias correction.

Adv: Demonstrated improvement in ETKF ensemble forecasts in NCAR model.

Dressing Samples



Dressed Ensemble



Dis: Only works if too little spread, not too much.

Ref: Roulston and Smith (Tellus, 55A, p 16); Wang and Bishop (QJRMS, 2005; picture above)

Bayesian model averaging (BMA)

$$p(y \mid f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k \; g_k(y \mid f_k)$$

Weighted sum of kernels centered around individual, bias-corrected forecasts.

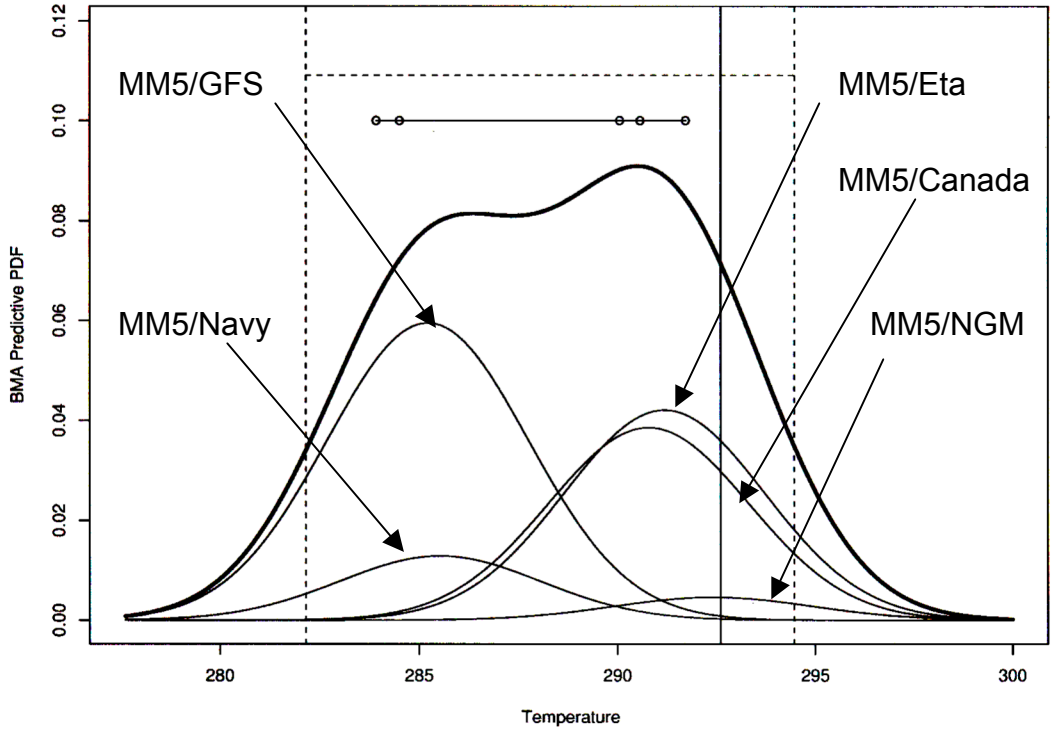


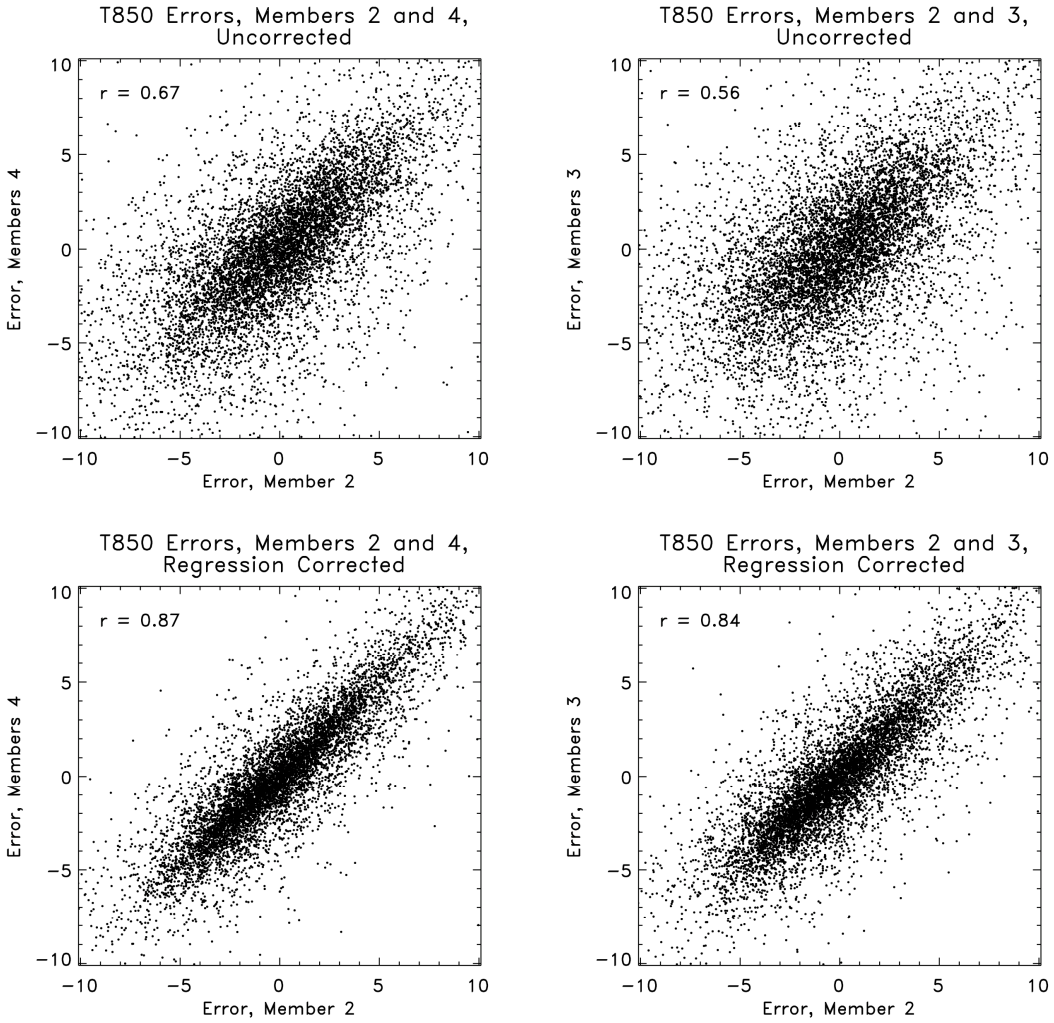
Figure 3: BMA predictive PDF (thick curve) and its five components (thin curves) for the 48-hour surface temperature forecast at Packwood, Wash., initialized at 0000 UTC on June 12, 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

Advantages: Theoretically appealing. No parameterized distribution assumed, weights applied proportional to their independent information (in concept).

Disadvantages: When trained with small sample, BMA radically de-weighted some members due to "overfitting" See Hamill, *MWR*, Dec. 2007.

Ref: Raftery et al., MWR, 2005. Wilson et al., MWR, 2007

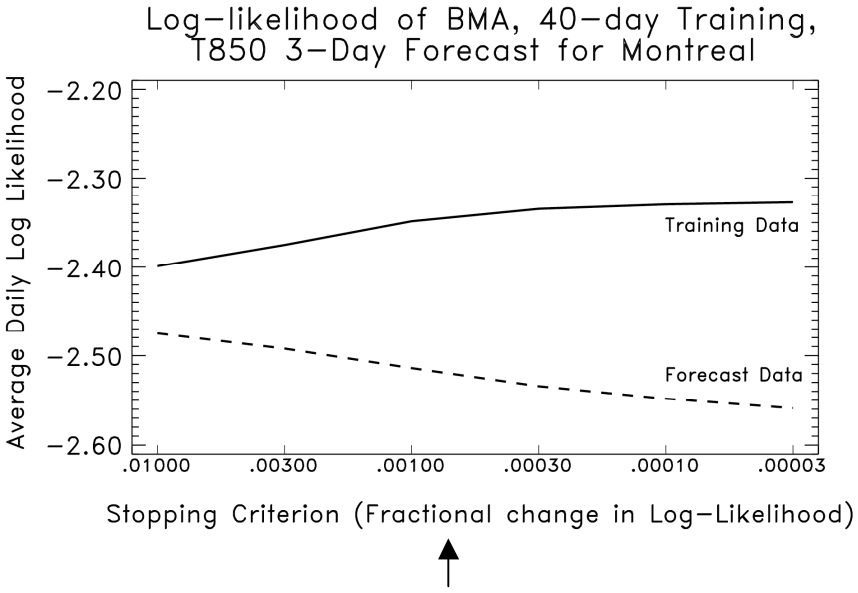
Why BMA's unequal weights? (1) regression correction accentuates error correlations.



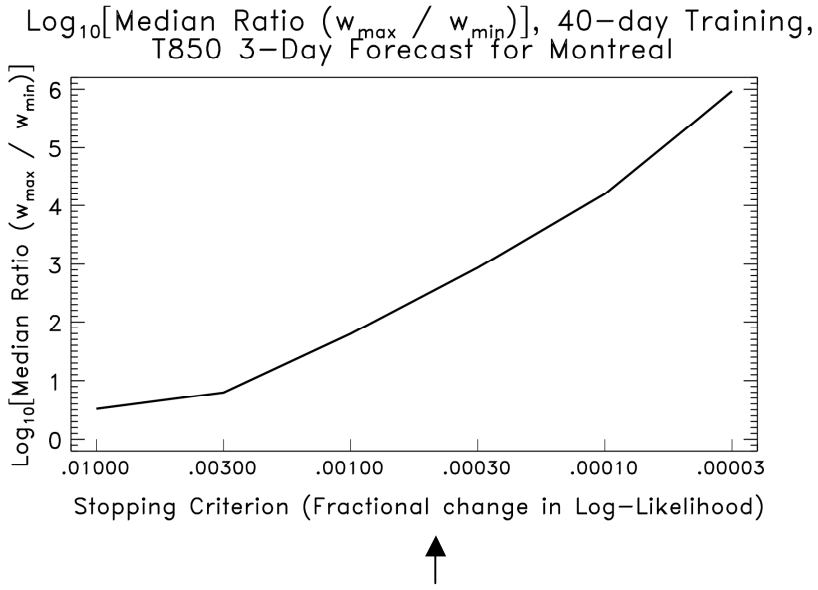
Ref: Hamill, *MWR*, Dec. 2007

Why BMA's unequal weights? (2) E-M overfits with little training data

An "estimation-minimization" (E-M) algorithm is used to determine the weights applied to ensemble members. If two forecasts have highly co-linear errors, E-M will weight one very highly, the other very little.

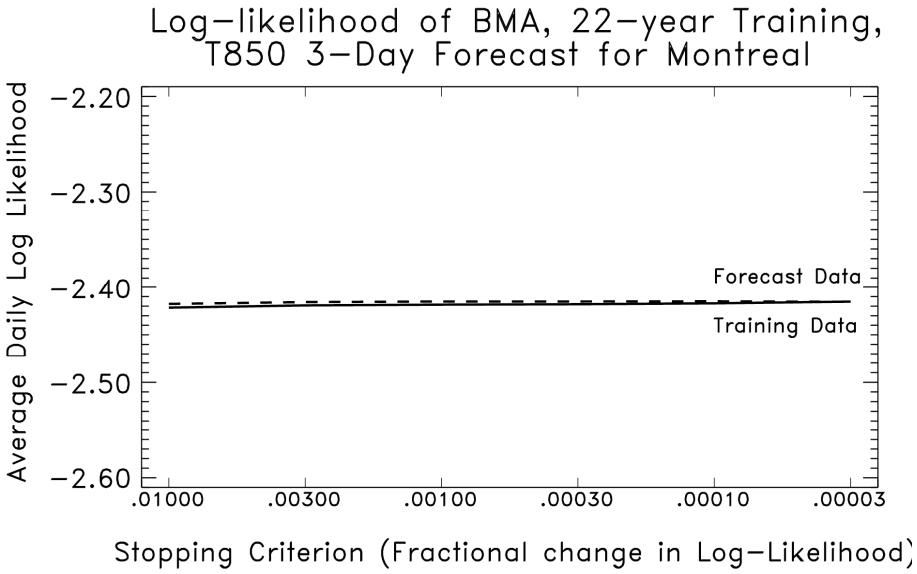


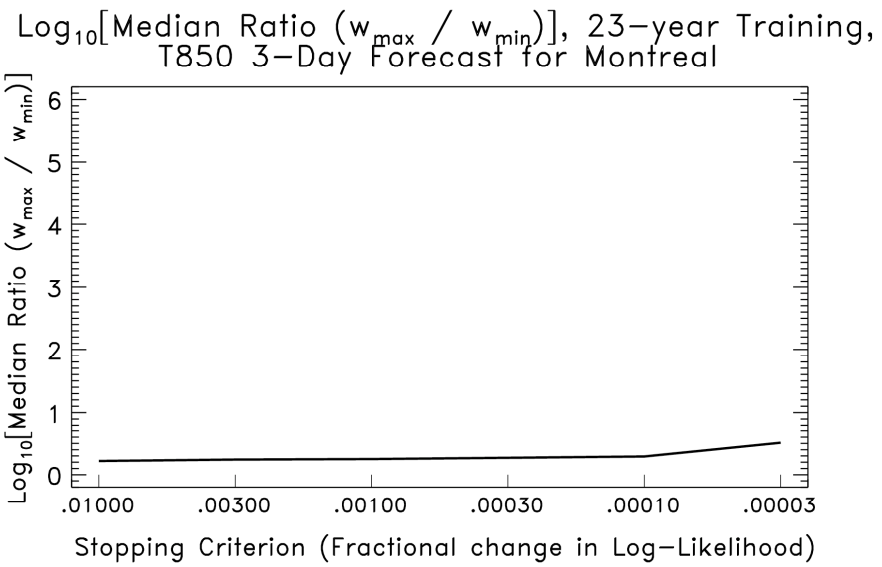
E-M is an iterative technique, and we can measure the accuracy of the fit to the data through the log-likelihood. Something odd happens here; as the E-M convergence criteria is tightened, the fit of the algorithm to independent data gets worse.



This plots the ratio of the weights of the highest-weighted member to the lowest-weighted member. As the convergence criterion is tightened, the method increasingly weights a few select members and de-weights others.

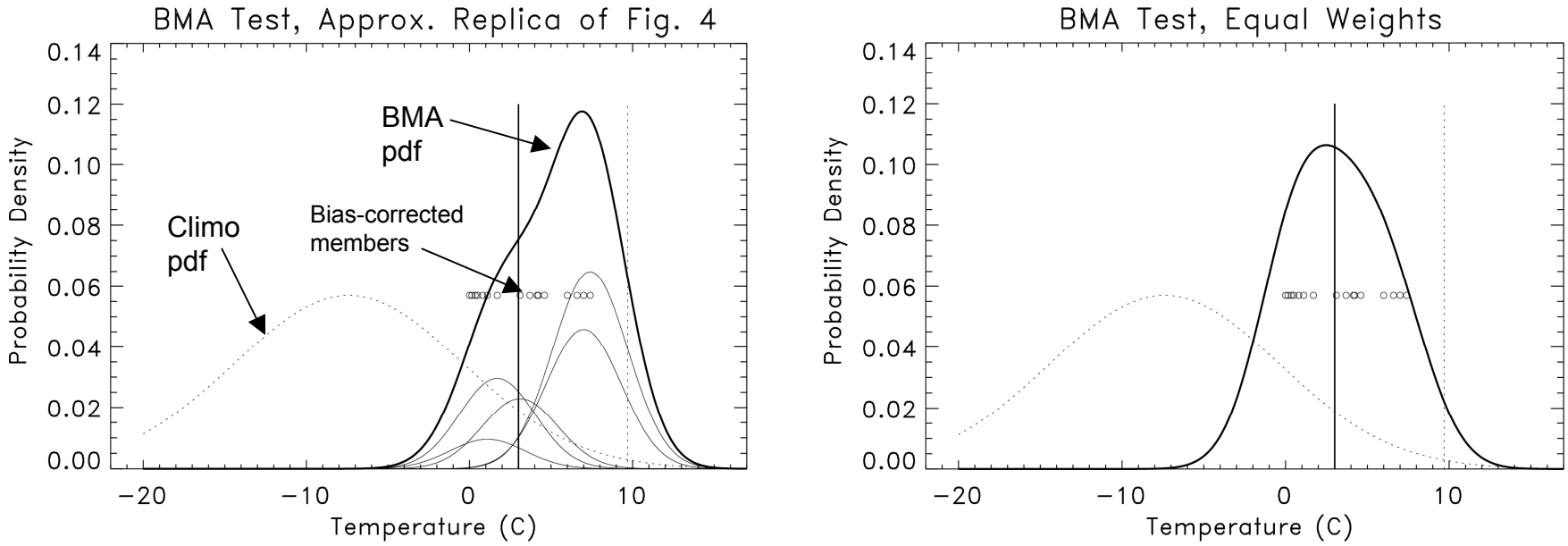
(BMA overfitting not a problem with 2+ decades training data)





With reforecast data set, we can train with a very large amount of data. When we do so, the weights applied to individual members are much more equal. This indicates that the unequal weighting previously is incorrect.

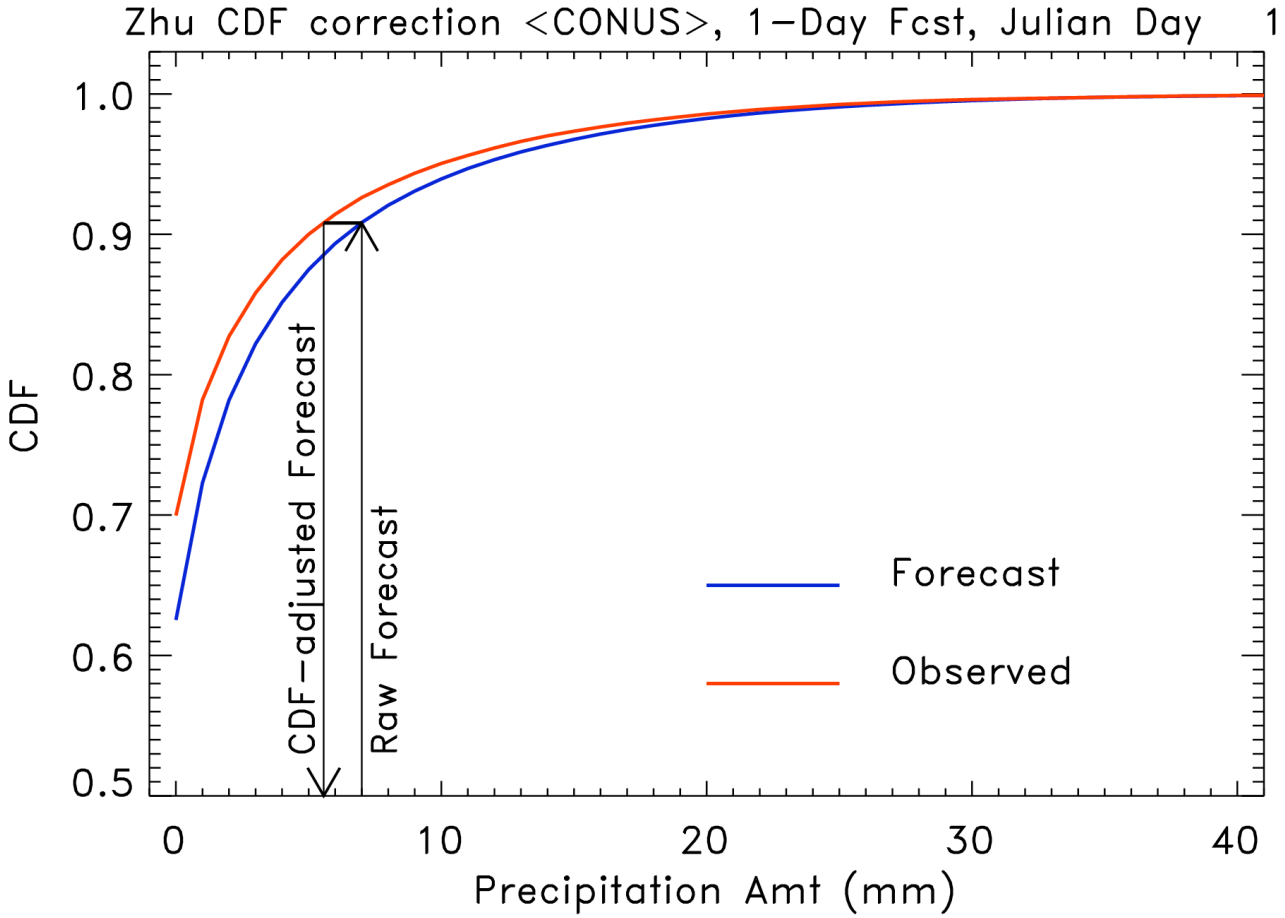
BMA's problem: an example



Here's a test of BMA in the winter season for a grid point near Montreal. BMA ends up highly weighting the warmest members (inappropriately so), thus producing a very high probability of a warm forecast.

Ref: Laurie Wilson, Met Service Canada, personal correspondence. See also upcoming Bishop et al. MWR manuscript.

Another problematic method: CDF-based corrections



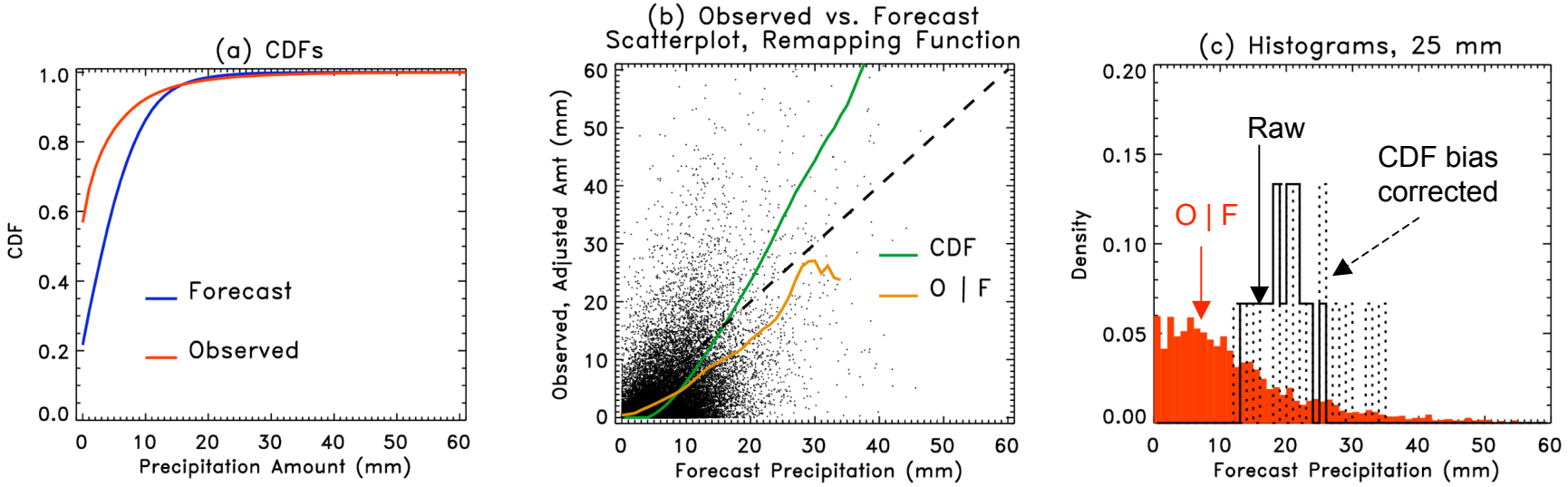
Use difference in CDFs to correct each ensemble member's forecast. In example shown, raw 7-mm forecast corrected to ~5.6 mm forecast.

NOTE: bias only, not spread correction or downscaling.

Ref: Zhu and Toth, 2005 AMS Annual Conf., and many others

CDF corrections: example of problem

1-day forecasts in Northern Mississippi (US), mid-August.
Consider a forecast precipitation of 25 mm.



CDF-based corrections at high amounts suggest further increasing precipitation amount forecast. O|F indicates decrease.

At root of problem is assumption that Corr (F,O) ≈1.0

Non-homogeneous Gaussian regression

- **Reference**: Gneiting et al., *MWR*, **133**, p. 1098
- **Predictors**: ensemble mean and ensemble spread
- **Output**: mean, spread of calibrated Gaussian distribution

$$f^{CAL}\left(\overline{\mathbf{x}}, \sigma\right) \sim N\left(a + b\overline{\mathbf{x}}, c + d\sigma\right)$$

- **Advantage**: leverages possible spread/skill relationship appropriately. Large spread/skill relationship, c ≈ 0.0, *d* ≈1.0. Small, *d* ≈ 0.0
- **Disadvantage**: iterative method, slow…no reason to bother (relative to using simple linear regression) if there's little or no spread/skill relationship.

Is there a “best” calibration technique?

Using Lorenz ‘96 toy model, direct model output (DMO), rank histogram technique, MOS applied to each member, dressing, logistic regression, non-homogeneous Gaussian regression (NGR), “forecast assimilation”, and Bayesian model averaging (with perturbed members assigned equal weights) were compared. Comparisons generally favored logistic regression and NGR, though differences were not dramatic, and results may not generalize to other forecast problems such as ones with non-Gaussian errors.

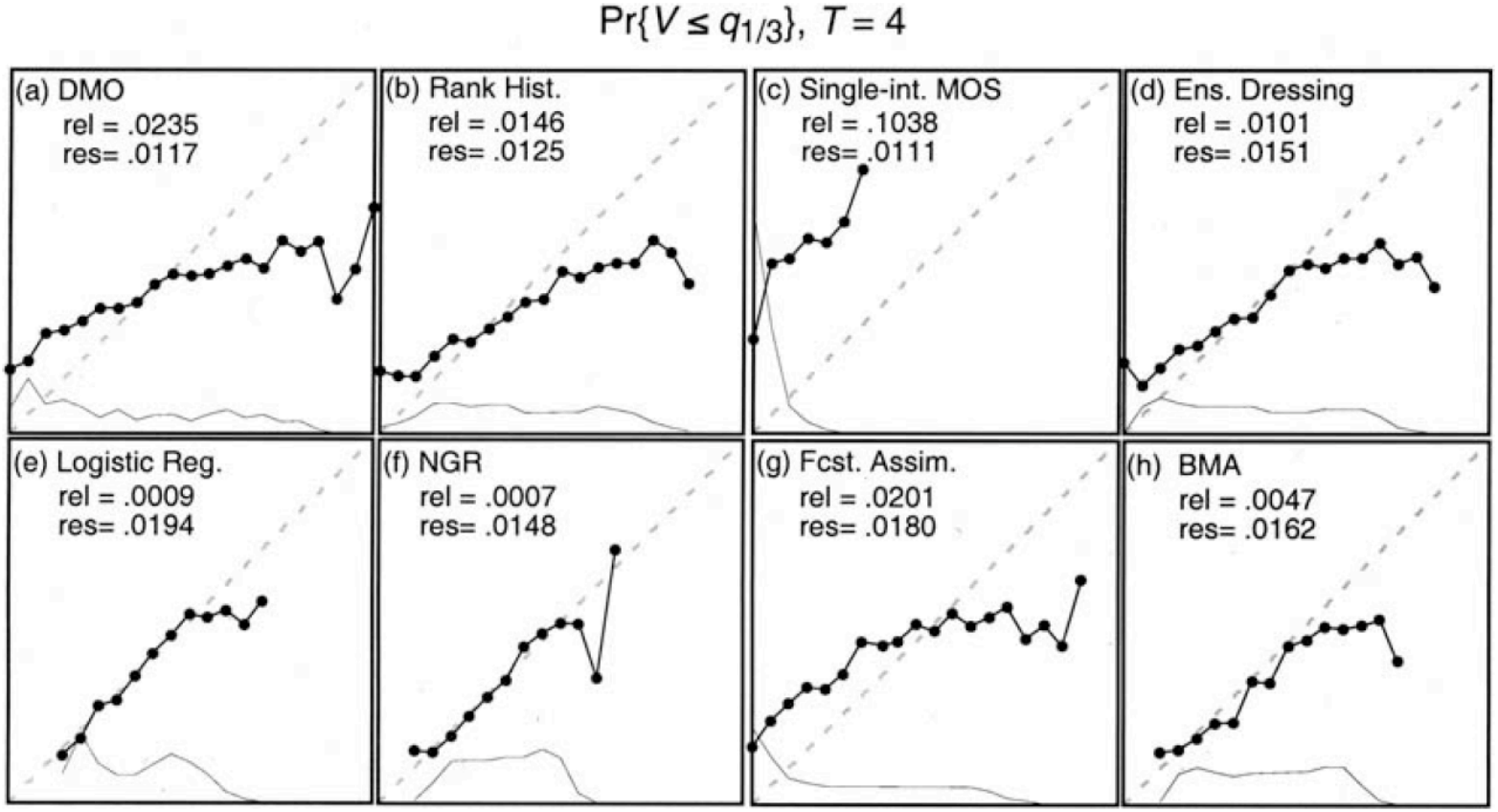
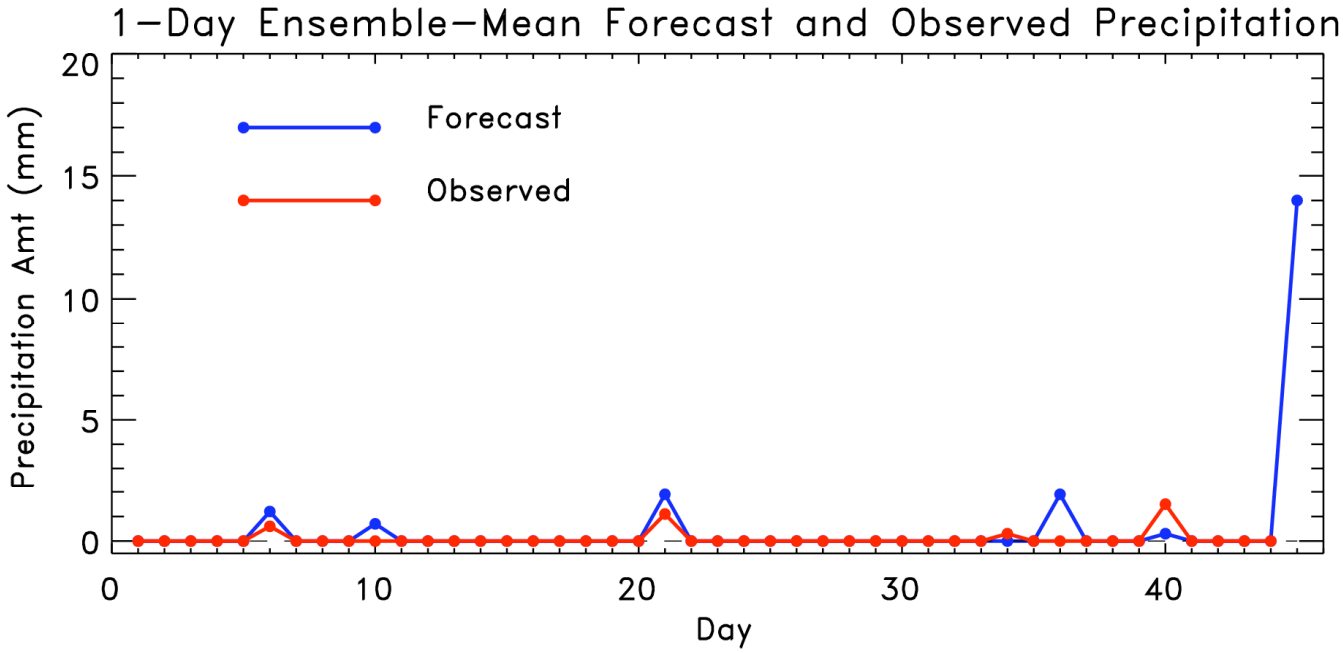


Figure 8. *As Figure 5, for Pr{V ≤ $q_{1/3}$} at lead time T = 4.*

Ref: Wilks, *Met. Apps*, 2006, **13**, p. 243

Calibration of PQPF & rare events: importance of sample size

Want lots of old forecast cases that were similar to today's forecast. Then the difference between the observed and forecast on those days can be used to calibrate today's forecast.



[More to say on this topic in reforecast seminar]

Combination:
multiple models, multiple parameterizations

- "One possible approach to operational mesoscale guidance is to produce an ensemble forecast using a combination of different initial conditions ***and different trigger functions***."

  – Dave Stensrud, 1994, from Stensrud and Fritsch, *MWR*, Sep. 1994 (part III).

Photo credit: Monty Python and the Holy Grail

Multi-(whatever) ensembles

- Potential plusses:
  - Provide different but still plausible predictions.
  - Models may have particular strong/weak aspects. Leverage the strong, discount the weak.
  - Implicitly samples analysis uncertainty through assimilation of somewhat different sets of observations, use of different data assimilation techniques.
  - Leverage each other's hard work and CPU cycles.
- Potential minuses
  - Models may all be developed under similar set of assumptions (e.g., which terms to neglect in equations of motion). What if some of these are consistently wrong and forecasts have similar biases?
  - Complex task to share data internationally in real time.
  - Must be flexible to use whatever is available, given outages / production delays.

Krishnamurti's "superensemble"

Table 1. The 850-hPa wind rms error (ms^{-1}) for 3-day prediction.

	ECMWF	RPN	UKMO	NCEP	NRL	BMRC	JMA	Ensemble mean	Superensemble
Globe	4.1	4.7	5.8	5.8	4.2	4.7	4.8	4.0	3.5
Tropics	2.7	3.5	3.4	4.5	3.1	3.4	3.5	2.7	2.2
Monsoon	2.6	3.4	2.9	4.6	3.1	3.4	3.5	2.7	2.0
Europe	2.0	2.2	2.9	3.3	2.2	2.2	2.0	2.0	1.7
United States	2.6	3.1	3.8	4.5	2.9	3.0	3.1	2.9	2.5
Northern Hemisphere	3.0	3.7	3.8	4.8	3.2	3.6	3.9	3.3	2.8
Southern Hemisphere	4.8	5.4	7.2	6.6	5.0	5.5	5.6	4.6	4.2

Table 2. Hurricane track rms errors (degrees).

Model	NHC	NOGAPS	UKMO	GFDL	FSU	Ensemble average	Cross validation	Superensemble
Day 1	1.5	1.7	1.6	1.7	1.7	1.2	1.2	0.9
Day 2	2.8	3.4	2.6	3.0	3.4	2.4	1.9	1.5
Day 3	3.4	3.8	3.9	5.5	4.8	2.6	2.6	1.9

Table 3. Hurricane intensity forecast rms errors (ms^{-1}).

Model	NHC	GFDL	FSU	Ensemble average	Cross validation	Superensemble
Day 1	6.6	10.0	11.0	6.5	5.7	5.1
Day 2	9.9	10.3	11.7	8.8	9.0	7.7
Day 3	14.0	12.1	14.5	12.3	12.0	9.6

Multi-model, multiple linear regression using a relatively short training data set of recent past forecasts.

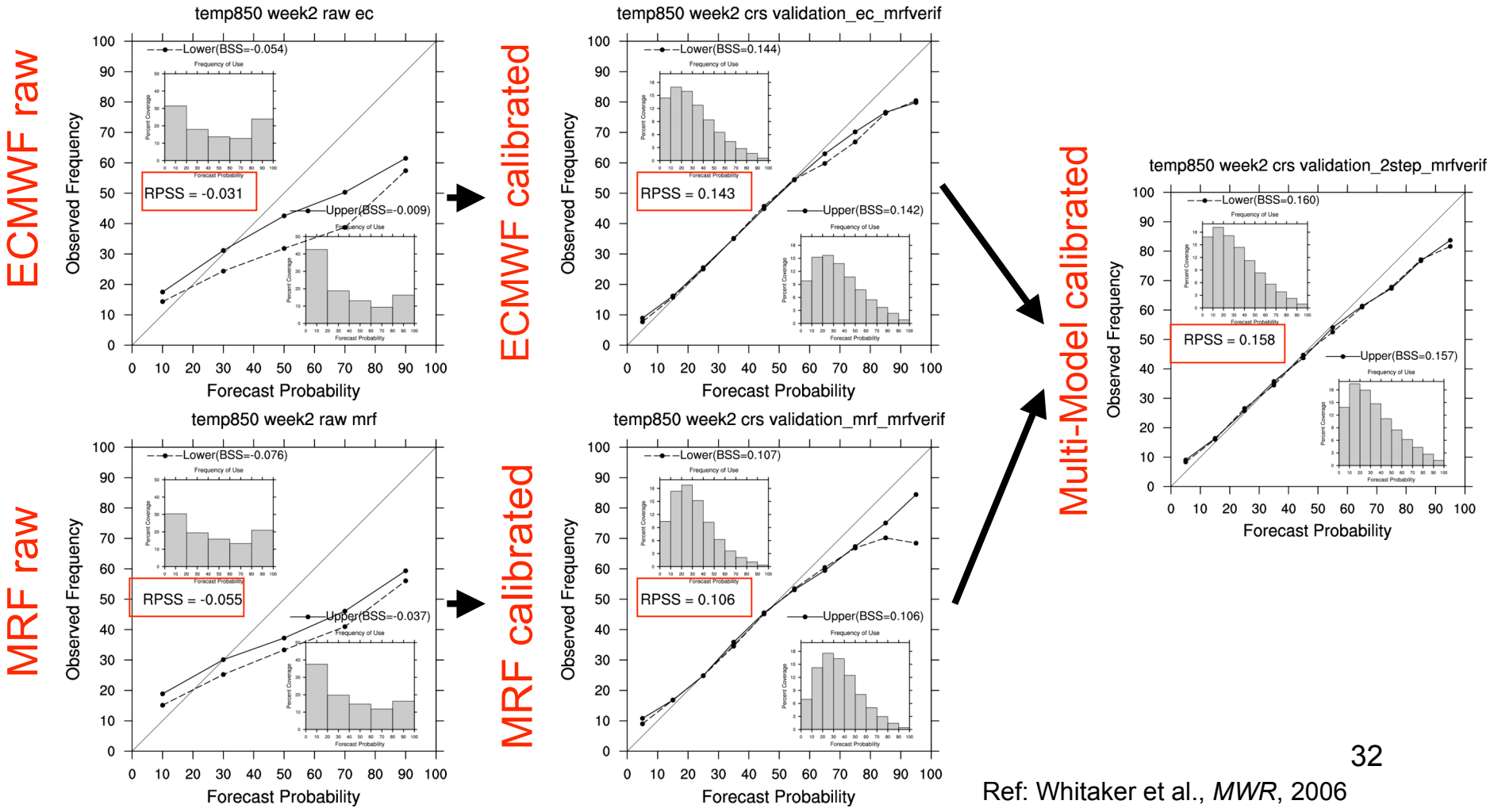
Despite use of the word "superensemble," the forecasts were expressed deterministically, though the regression analysis implicitly provides enough information to make probabilistic forecasts.

See also Vislocky and Fritsch, *BAMS*, July 1995.

Ref: Krishnamurti et al., *Science*, 3 Sep 1999

Combination + calibration example

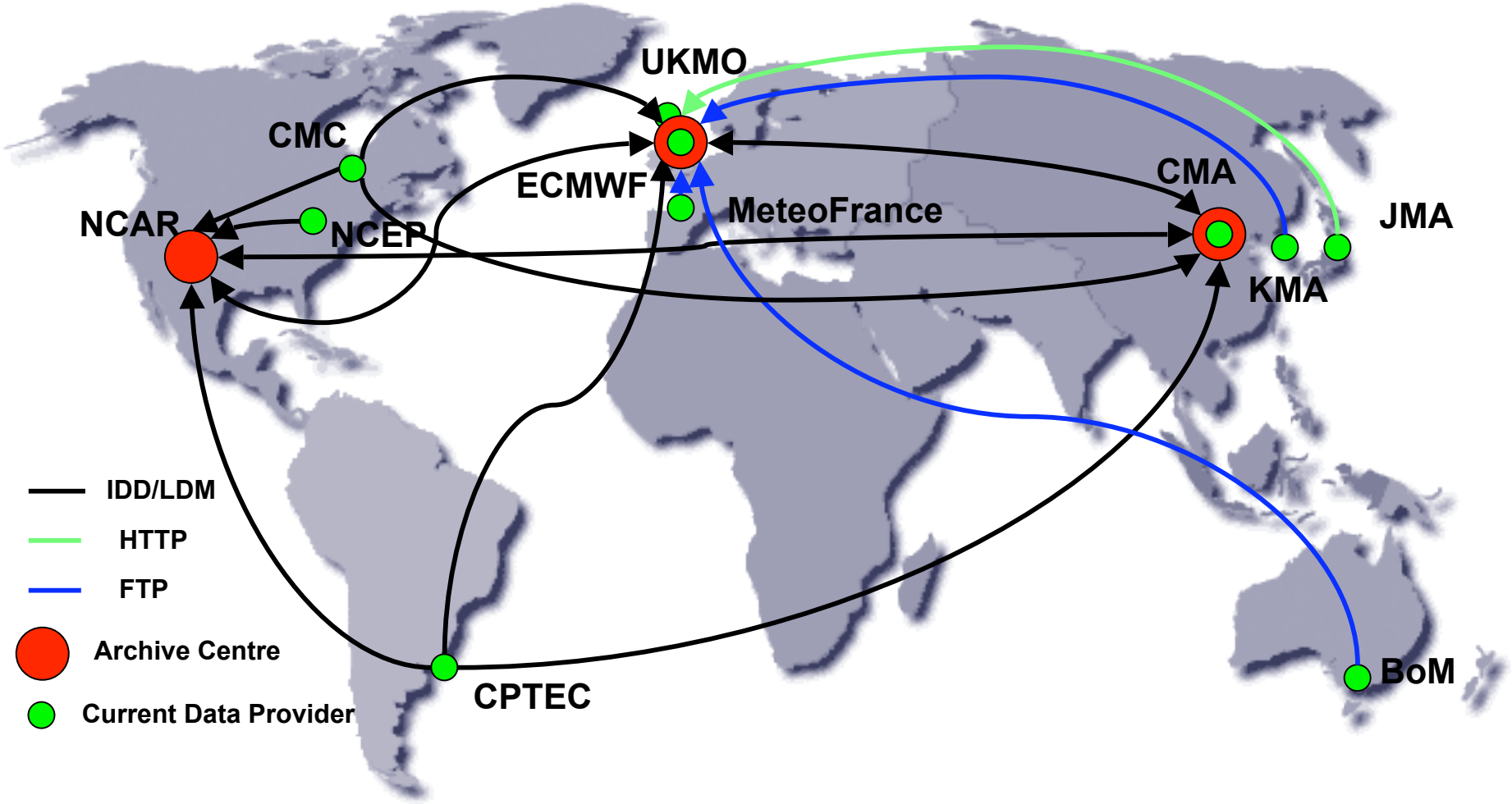
ECMWF produced 5-member reforecasts once every 2 weeks for 10 years
in DJF. Apply logistic regression to ECMWF, MRF (i.e. GFS), and both for week 2 terciles.



Ref: Whitaker et al., *MWR*, 2006

THORPEX Interactive Grand Global Ensemble (TIGGE)

archive centers and data providers



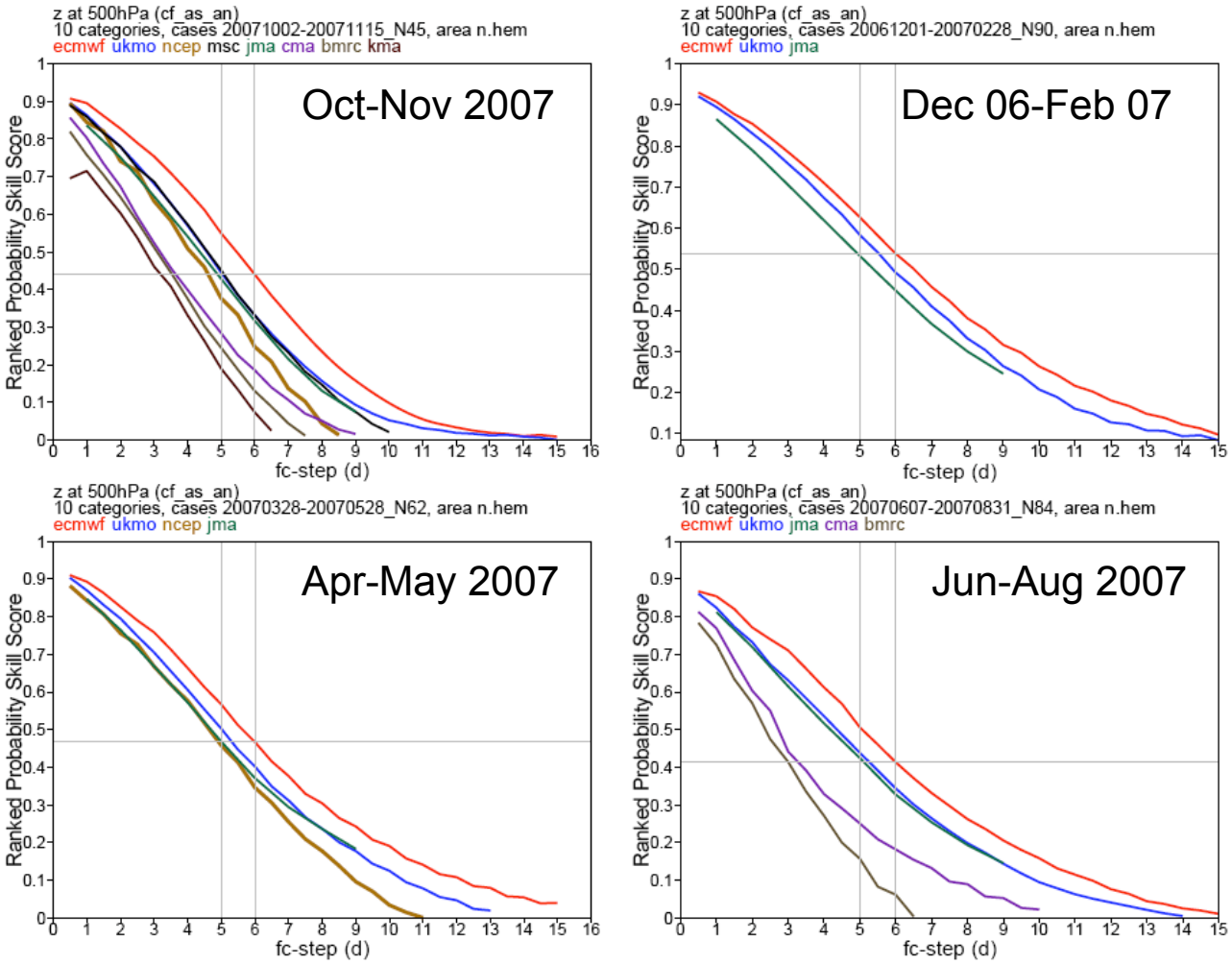


Figure 8: Average ranked probability skill score for the probabilistic prediction of Z500 over NH of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles, each verified against its own analysis, for four periods (due to data availability, not all forecasts were available for all periods):

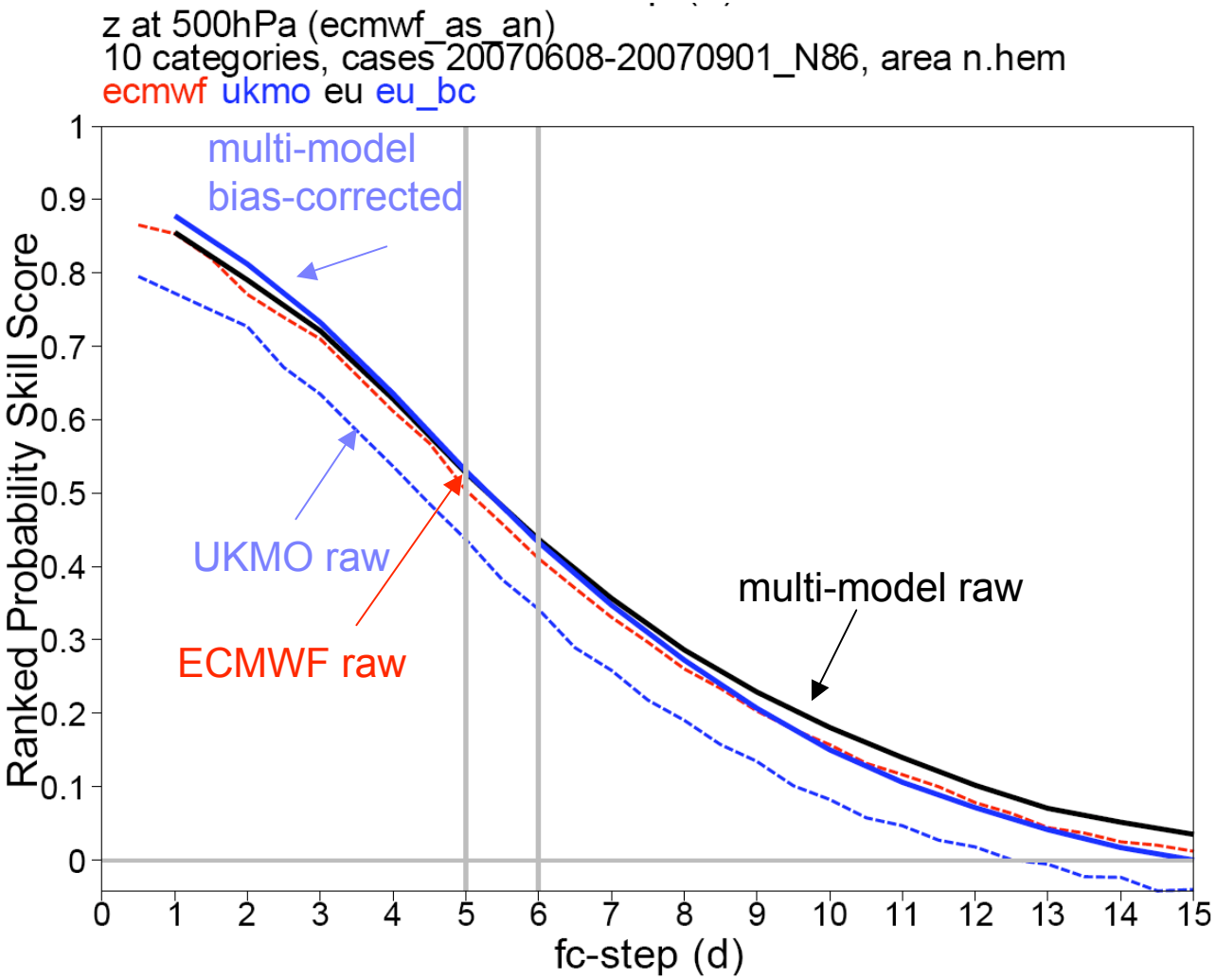
a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA
b: DJF07 (90 cases), EC, UKMO and JMA
c: AM07 (62 cases), EC, UKMO, NCEP and JMA
d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC

TIGGE, RPSS of Z500

- Skill of forecasts against own analyses for 4 different periods using TIGGE data
- Models obviously of different quality.

Ref: Park et al., 2008, ECMWF TM 548

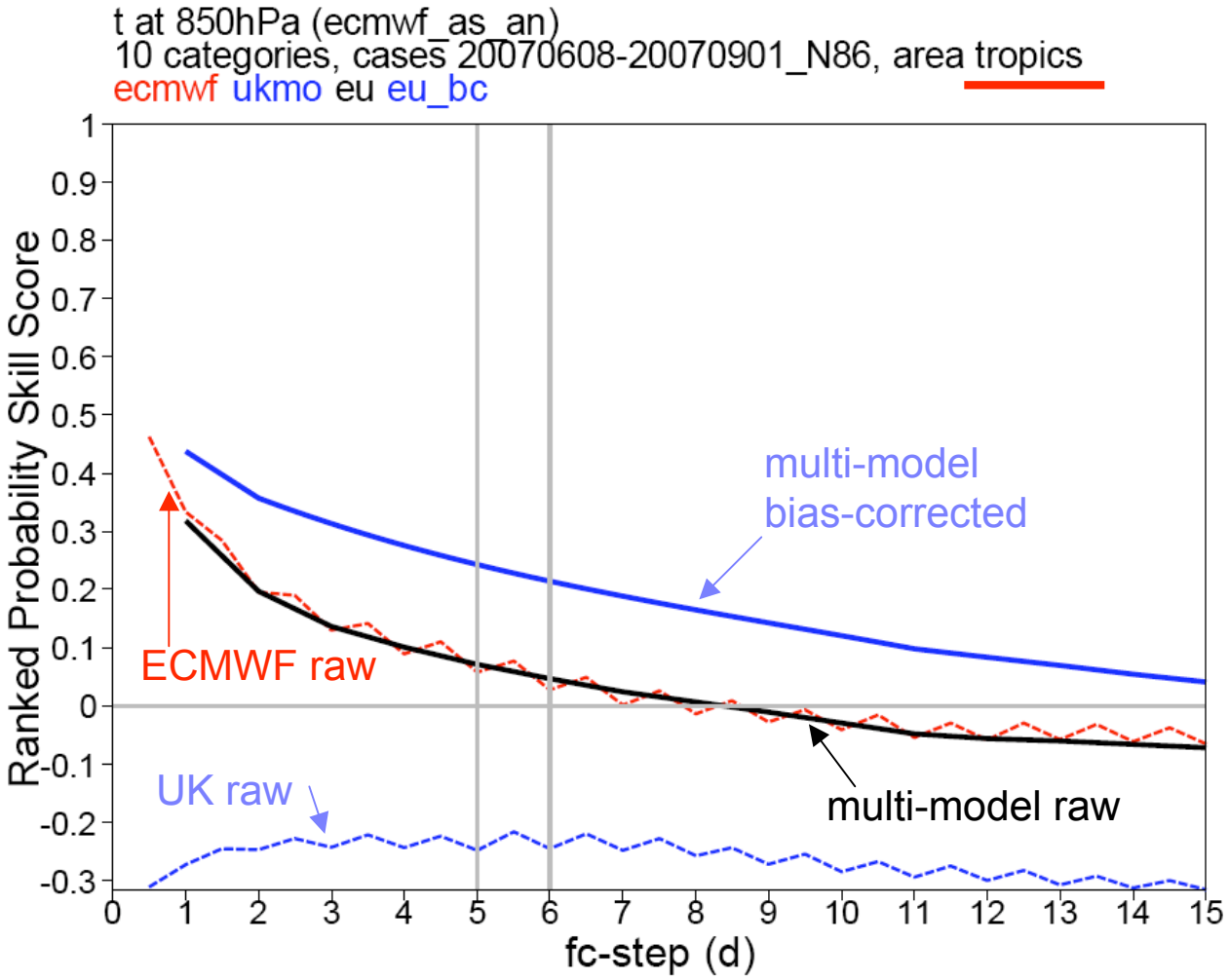
TIGGE, Z500, ECMWF & UKMO



- Trained with bias correction using last 30 days of forecasts and analyses.
- ECMWF analysis as reference.
- Conclusions:
  - (1) Small benefit from multi-model relative to best model.
  - (2) + impact of bias correction at short leads, - at long leads. [Reforecast seminar will discuss why]

Ref: Park et al., 2008, ECMWF TM 548

T850, two-model



- Trained with bias correction using last 30 days of forecasts and analyses.
- ECMWF analysis as reference.
- Conclusions:
  (1) UK so contaminated by systematic errors that its raw forecasts add no value.
  (2) Multi-model calibrated uniformly beneficial (presumably because of large, ~consistent biases)

Ref: Park et al., 2008, ECMWF TM 548

Example: flood events in Romania, October 2007

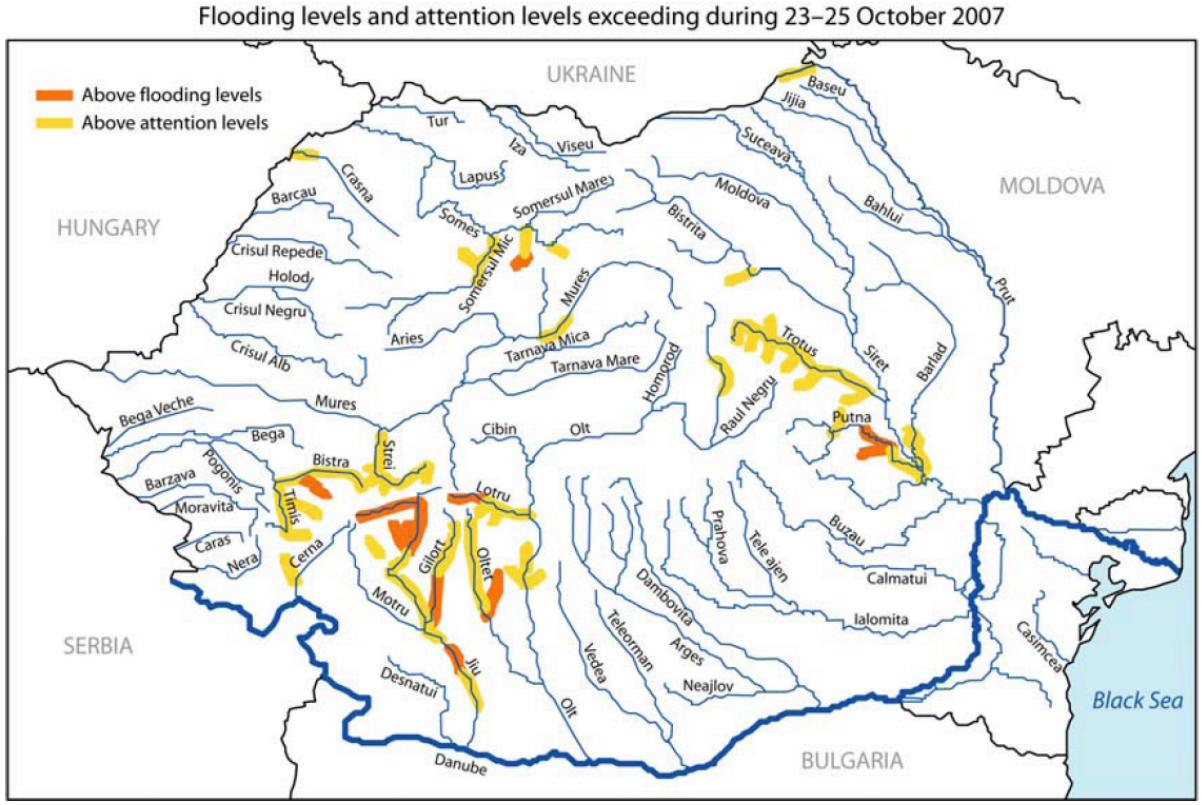
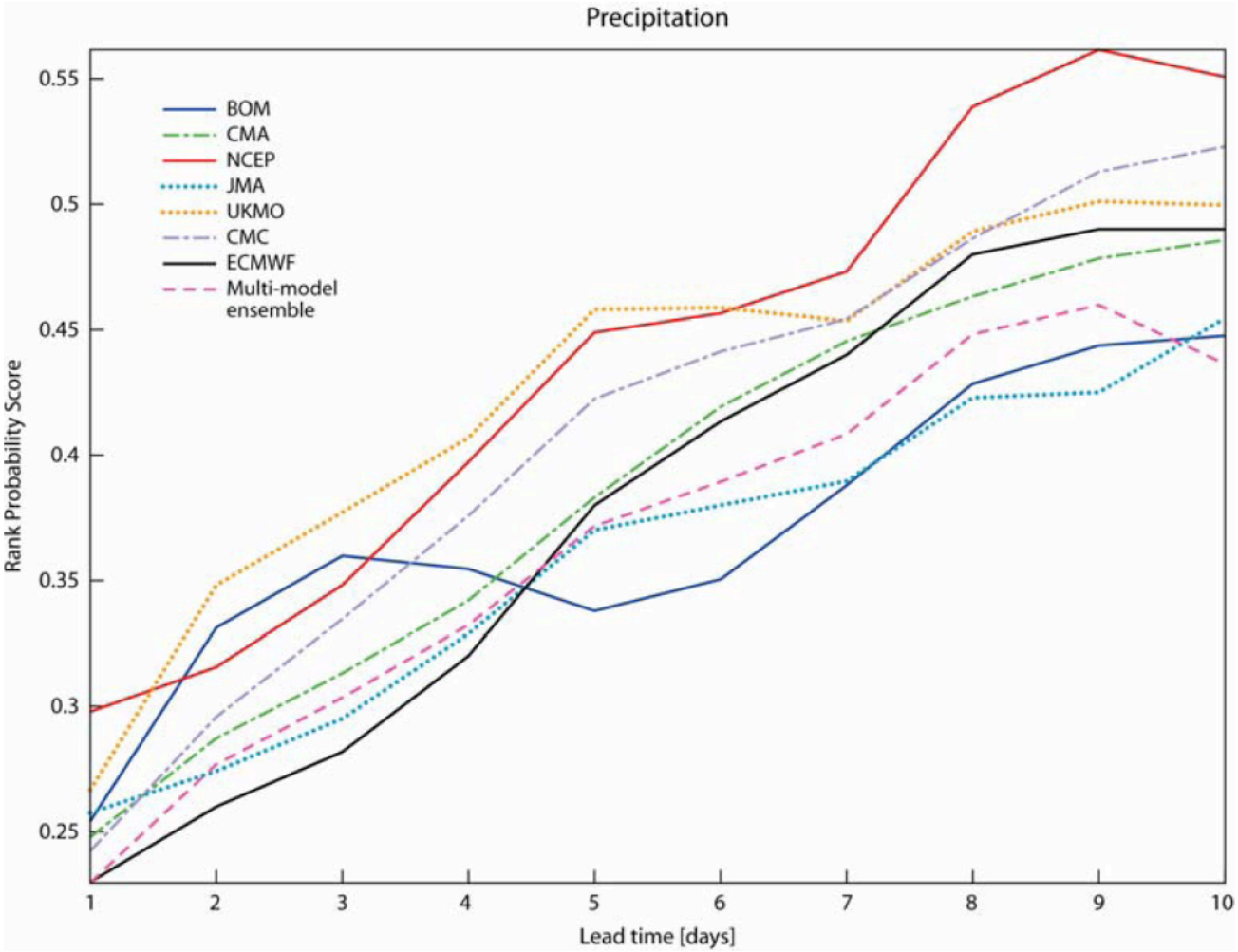


Figure 1: Flood Levels and Attention Levels in Romania exceeded during the 23rd to 25th of October 2007. The basemap for this figure has been kindly provided by E. Anghel from the National Institute of Hydrology and Water Management, Bucharest Romania.

- Drive routing models with TIGGE data.
- Verification over all basins…only a few flood.

Ref: Pappenberger et al., 2008, ECMWF TM 557

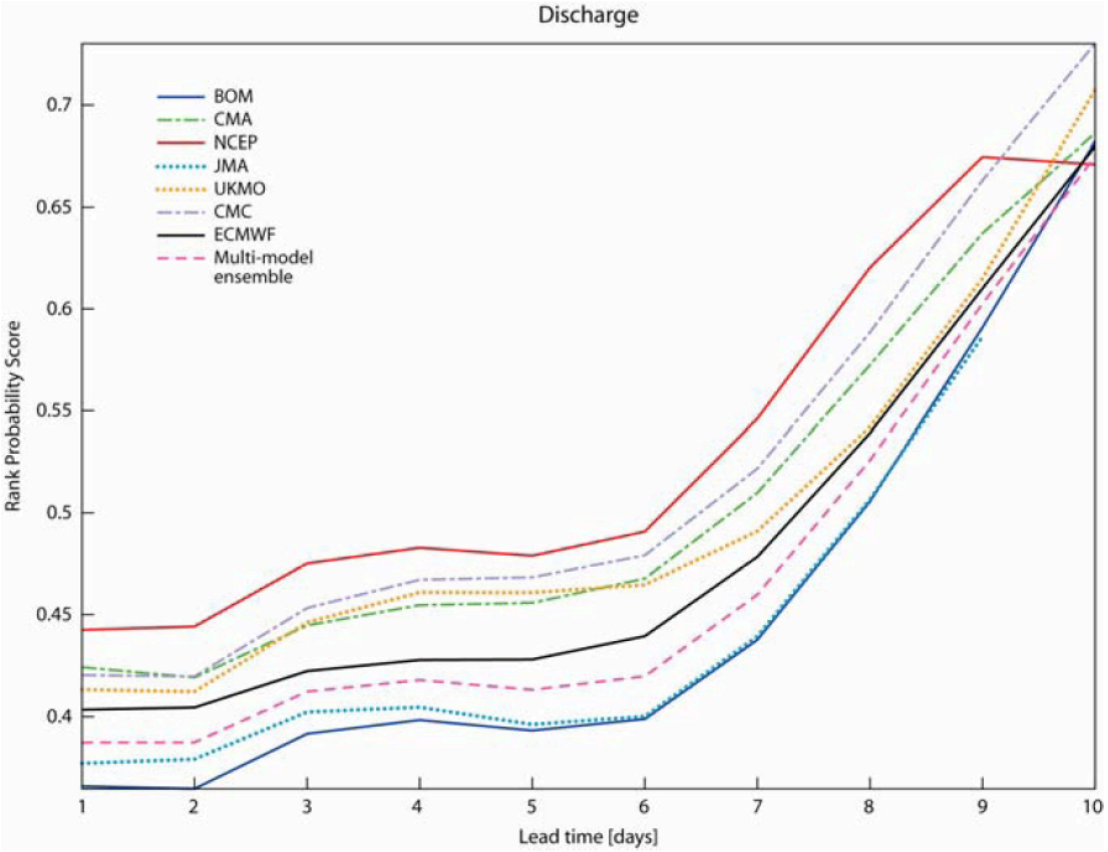
Precipitation forecast skill for a flash flood event in Romania



- Thresholds for RPS are 10th, 90th percentiles of observed distribution.
- ECMWF the best at short leads; JMA/BOM/multi-model the best at longer leads.

Ref: Pappenberger et al, 2008, ECMWF TM 557

Skill of streamflow forecasts



- Somewhat surprisingly, BOM and JMA forecasts scored very well in this regard, when considering both flood and non-flood events. This was because the other models had significant light precipitation over-forecast bias. Interesting difference between this and skill of precipitation forecasts themselves. This highlights how a chain of models can be nonlinearly affected by deficiencies in an earlier model.

Ref: Pappenberger et al, 2008, ECMWF TM 557

Some opinions on calibration and combination

- What's the best calibration technique?
  - Enlarging training sample size has a bigger effect for many variables than changing the calibration technique (see reforecast seminar).
  - Preferred techniques may vary from user to user. Hydrologists want bias-corrected and downscaled members, others want smooth pdfs.
  - KISS (Keep it simple, stupid). Increasing focus on non-parametric techniques that permit complex, multi-modal distributions. Often parametric distributions work just fine.

- Combination:
  - Good: multi-model ensemble.
  - Better: calibrated multi-model products using short training data sets, or calibrated single-model based on reforecasts.
  - Best: calibrated multi-model products based on reforecasts.